Theoretical Machine Learning - COS 511

Homework Assignment 1

Due Date: 22 Feb 2017, till 22:00

(1) Consulting other students from this course is allowed. In this case - clearly state whom you consulted with for each problem separately.

(2) Searching the internet or literature for solutions, other than the course lecture notes, is NOT allowed.

(3) All problems are weighted equally at 10 points each. Indicate on your problem set which four problems you choose to solve. Feel free to write down solutions for the other two as well, but your homework grade will only depend upon the four you mark to be graded.

Ex. 1:

Let $X = \mathbb{R}^2$ be the domain and $Y = \{0, 1\}$ be the label set of a learning problem. Let $\mathcal{H} = \{h_r\ ,\ r \in \mathbb{R}_+\}$ be a set of hypothesis corresponding to all concentric circles on the plane that classify as

$$h_r(x) = \begin{cases} 1 & \|x\|_2 \leq r \\ \\ 0 & o/w \end{cases}$$

Prove that under the realizability assumption $\mathcal{H}$ is PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) = O\left(\frac{\log \frac{1}{\delta}}{\varepsilon}\right)$$

Ex. 2: [agnostic means noise-tolerance]

Let $\mathcal{C}$ be an (agnostic) PAC learnable class. For each concept $h\mathcal{C}$, consider the concept $\hat{h}$

which is obtained by replacing the label associated with each domain entry $x \in X$ randomly with probability ε_0 every time x is sampled independently. That is:

$$\hat{h}(x) = \begin{cases} 1 & w.p. \ \frac{\varepsilon_0}{2} \\ 0 & w.p. \ \frac{\varepsilon_0}{2} \\ h(x) & o/w \end{cases}$$

Prove that there is an algorithm A that can ε-approximate any concept $\hat{h}$: that is, show that A can produce a hypothesis h_A that has error

$$\text{err}(h_A) \leq \frac{1}{2}\varepsilon_0 + \varepsilon$$

with probability at least $1 - \delta$ for every ε, δ with sample complexity polynomial in $\frac{1}{\varepsilon}, \log \frac{1}{\delta}$.

Ex. 3: [Proving Chernoff's bound]
In this exercise we'll prove **Chernoff's inequality**:
Let $x_1, x_2 ... x_k$ be independent random variables, each receiving the values $\{-1, 1\}$ w.p $\frac{1}{2}$. Define: $X = \sum_{i=1}^{k} x_i$, then for any real number $t > 0$:

$$\mathbb{P}[X \geq t] \leq e^{\frac{-t^2}{2k}}$$

(1) For the random variable X above, use Markov's inequality to show that for every $\lambda \geq 0$,

$$\Pr[X \geq t] = \Pr[e^{\lambda X} \geq e^{\lambda t}] \leq e^{-\lambda t} \cdot \prod_{i=1}^{k} \mathbf{E}[e^{\lambda x_i}] = e^{-\lambda t} \cdot (\frac{e^{\lambda}}{2} + \frac{e^{-\lambda}}{2})^k$$

(2) Prove that for all $\lambda > 0$, $(\frac{e^{\lambda}}{2} + \frac{e^{-\lambda}}{2}) \leq e^{\frac{\lambda^2}{2}}$ (hint: think of Taylor's theorem)

(3) Show how to conclude with the statement: $\mathbb{P}[X \geq t] \leq e^{\frac{-t^2}{2k}}$

Ex. 4:

For this problem, you need not be concerned about algorithmic efficiency.

- Suppose that the domain X is finite. Prove or disprove the following statement: If a concept h is PAC learnable by and algorithm with an hypothesis class $\mathcal{H}$, then $h \in \mathcal{H}$[1].
- Repeat the first part without the assumption that X is finite. In other words, for the case that the domain X is arbitary and not necessarily finite, prove or disprove that if h is PAC learnable by $\mathcal{H}$, then $h \in H$.

Ex. 5:

Extend the no free lunch theorem to state the following:
There exists a domain X such that for all $\varepsilon > 0$, for any integer $m \in N$, learning algorithm A which given a sample S produces hypothesis h_S^A, there exists a distribution D and a concept $h : X \mapsto \{0, 1\}$ such that

- $\text{err}_D(h) = 0$
- $\mathbf{E}_{S \sim D^m}[\text{err}(h_S^A)] \geq \frac{1}{2} - \varepsilon$

Ex. 6

This exercise illustrates the fact that VC-dimension of hypothesis class $\mathcal{H}$ does not necessarily depend on the number of free parameters used to characterize $\mathcal{H}$. Consider the hypothesis family of sine functions for binary classification $\{t \to \sin(\omega t) : \omega \in \mathbb{R}\}$. A point is is labeled 1 if above the sine curve, and 0 if below it.

(1) Show that for any $x \in \mathbb{R}$ the points $x, 2x, 3x, 4x$ cannot be shattered by this family of sine functions.
(2) Show that the VC-dimension of the family of sine functions is infinite (Hint: show that $\{2^{-m} : m \in \mathbb{N}\}$ can be fully shattered for any $m > 0$).

[1]To prove the statement, you of course need to give a proof showing that it is always true. To disprove the statement, you can simply provide a counterexample showing that it is not true in general.